Comment on ‘‘On-Line Gibbs Learning’’

Biehl, M; Riegler, P

Published in:
Physical Review Letters

DOI:
10.1103/PhysRevLett.78.4305

Comment on "On-Line Gibbs Learning"

On-line learning in feed-forward neural networks has attracted considerable attention among physicists in recent years. Apart from the possibility to study exactly the dynamics of known standard algorithms, statistical physics provides a workshop in which to develop efficient and even optimal training schemes from first principles (see, e.g., [1], and references therein).

In [2], Kim and Sompolinsky give an interpretation of on-line learning in the context of equilibrium statistical mechanics. We focus in this Comment on their results concerning the learning of nonsmooth problems with networks of threshold units, in which according to [2] the main novelty of the on-line Gibbs algorithm lies. The authors claim that their (p. 3024), third paragraph) "on-line Gibbs algorithm is the first on-line algorithm that guarantees convergence to the minimal generalization error." Results from optimal on-line learning (e.g., [1]) lead to the conclusion that this cannot be true. We demonstrate in the following, that, indeed, the on-line Gibbs algorithm fails to converge to the (locally) minimal value of ϵ_g if its parameter λ is chosen improperly.

As a specific example we have studied the case of the simple perceptron in the presence of uniform output noise (example (2) in [2]). We have solved this model exactly in the limit $N \to \infty$ (analogous to Refs. [12,13] in [2]) and find that not only the appropriate power law of the time dependence ($\lambda = \lambda_0 \alpha^2$) has to be guessed correctly but also the free parameter λ_0. An analysis of the fixed point properties of the corresponding on-line dynamics shows that λ_0 has to be smaller than a critical value which depends on the, in general, unknown noise level p. Otherwise the algorithm does not converge to the minimal value of ϵ_g. In particular, there is no value of λ_0 that guarantees convergence to the minimum generalization error for *all* noise levels (cf. Fig. 1), in contrast to the above cited conclusion of the Letter. Above the p-dependent critical value of λ_0 the system asymptotically approaches a state with a nonzero value of $\epsilon_g - \epsilon_{\min}$ (note that local minima do not exist in this model).

The authors state in [2] that, indeed, the [p. 3023, after Eq. (12)] "coefficient λ_0 has to be smaller than some system-dependent cutoff value." However, this "system dependence" is truly on unknown properties of the training data (here: the noise level p). Furthermore, the consequence of a mismatched choice of λ_0 is not just a suboptimal convergence rate, but a residual error $\epsilon_g > \epsilon_{\min}$.

This specific example demonstrates the general necessity of knowledge of some problem inherent properties and consequently it suggests to devise estimation schemes for such quantities. The constructive nature of the optimization procedure (as applied in, e.g., [1,3]) points out the relevant features necessary for efficient and optimal learning. Whether an arbitrarily constructed *ad hoc* algorithm performs satisfactorily relies on how well it reproduces the set of relevant characteristics.

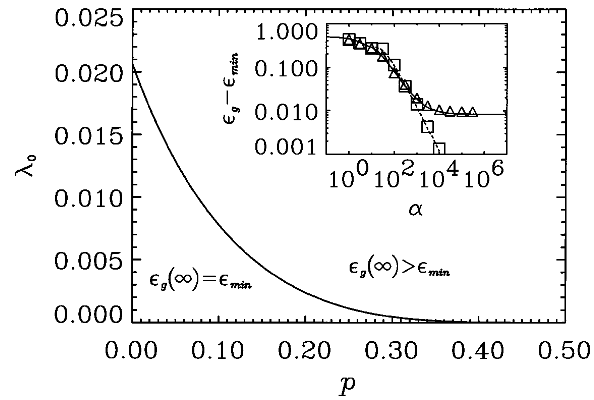

FIG. 1. Critical value of λ_0 (as explained in the text) vs the output noise level p for the perceptron. For values below the solid line the system converges to $\epsilon_{\min} = p$, above the line ϵ_g saturates at a higher value. The inset shows the learning curves for $N \to \infty$ and $p = 0.2$ with a subcritical $\lambda_0 = 0.001$ (dashed) and a mismatched value $\lambda_0 = 0.005$ (solid). Symbols display the results of simulations ($N = 100$, single runs).

In a sense, the specific algorithms for nonsmooth functions discussed in [2] provide very good approximations to the optimal ones and thus yield similar asymptotic behavior. Also, the critical dependence on the noise level, emphasized in this Comment, is in complete analogy with the findings of [1] and Ref. [13] in [2]. However, the on-line Gibbs algorithm lacks the important estimation of certain problem inherent quantities, as, for instance, λ_0 or its true physical equivalent, the noise level p.

Another example for such a problem inherent quantity appears in the context of a time dependent rule. The on-line Gibbs algorithm does not exploit the actual performance (ϵ_g) which would be necessary in order to cope with a rule that changes in time. This has already been shown in [3] and Ref. [5] of [2].

In conclusion, the success of on-line Gibbs learning is not guaranteed but depends critically on a certain knowledge of the nature of the learning task and its parameters, just as it does for any other practical algorithm.

M. Biehl and P. Riegler
Institut für Theoretische Physik
Universität Würzburg
D-97074 Würzburg, Germany

Received 16 October 1996 [S0031-9007(97)03204-3]
PACS numbers: 87.10.+e

[1] M. Copelli *et al.,* Phys. Rev. E **53**, 6341 (1996).
[2] J.W. Kim and H. Sompolinsky, Phys. Rev. Lett. **76**, 3021 (1996).
[3] O. Kinouchi and N. Caticha, J. Phys. A **26**, 6161 (1993).

